<CH>How to perform a critical analysis of a randomized controlled trial

<AU>Candice Estellat MD,* David J. Torgerson[1], PhDPhilippe Ravaud MD, PhD

<ADD>INSERM U738, Paris, France, [1]Dpt of Health Sciences, University of York UK

<CORR>*Corresponding author. INSERM U738, Department d'Epidémiologie, de Biostatistique et de Recherche Clinique, Groupe Hospitalier Bichat -- Claude Bernard, 46 rue Henri Huchard; 75877 Paris Cedex 18, France

<CORR>Tel.: 33 (0)1 40 25 79 44; fax 33 (0)1 40 25 67 73

<CORR>E-mail address: candice.estellat@bch.aphp.fr

<ABS>Abstract

Given the large amount of medical literature of varying methodological quality, care must be taken when translating the results of randomized controlled trials into clinical practice. To assist in this translation process, we provide a method that involves answering three main questions: 'Can I trust the results?' 'How do I understand the results?' and 'To whom do the results apply?' To answer the first question, we describe important items that help in judging the reliability of the findings. For the second question, we address the clinical and statistical significance of results by looking at the size and precision of the effect. Finally, we raise the issue of external validity and reporting adverse effects to determine which patients may best benefit from the new intervention.

<KEY>**Key words**: Bias; Evidence-based medicine;Randomization; Randomized controlled trial; Sample size; Selection bias

Given the large amount of medical literature of varying methodological quality, one must be careful when transposing the results of a randomized controlled trial (RCT) to clinical practice. Here, we give some keys to assess the reliability and understand the results of such trials. Because of the wide-ranging fields of medical literature, we focus on two-group parallel RCTs designed to demonstrate whether one intervention is better than

another. Methodological issues are universal, but as much as possible we try to give examples in the field of rheumatology.

We base our approach to the critical analysis of results of RCTs for translation to clinical practice on the systematic approach that Guyatt et al. proposed for critical analysis of medical literature.1,2 Our approach is based on three main questions: ‘Can I trust the results?’, ‘How do I understand the results?’ and ‘To whom do the results apply?’. To help answer the first question, we explain important items that help in judging the reliability and validity of the findings. For the second question, we address the clinical and statistical significance of results by looking at the size and precision of the effect. Finally, we raise the problem of external validity and reporting of adverse effects to determine which patients may best benefit from the new intervention. A minimal set of items that must be reported to allow for critical analysis of results of an RCT are in the CONSORT statement checklist,3 which could be useful for both authors and readers.

<A>Can I trust the results?

To evaluate the effect of a treatment, as in many experiments, the best way is to compare two groups -- one undergoing the treatment and one not (control group) -- to see which has the best outcome. However, for this comparison to be valid and to avoid bias, the groups must be similar at baseline, undergo the same care apart from the treatment under study and be assessed in the same way at the end of the study. A bias is anything that erroneously influences the difference between groups. Specific methods developed to overcome sources of bias in clinical trials are summarized in Figure 1, which is adapted from Greenhalgh,4 and detailed below.

<Figure 1 here>

<B>Were patients randomized and was the randomization concealed?

Randomizing patients between the treatment and control groups increases the probability of obtaining similar patients in the two groups at baseline and thus avoids selection bias. Random allocation is the only way to ensure no systematic differences between intervention groups in known but also unknown factors that might affect the outcome.

Proper randomization rests on two equally important elements: generation of the sequence and its implementation.

Adequate generation of the sequence usually involves a random-number table or a computerized random-number generator. Deterministic allocation methods (sometimes called ‘quasi-random’ methods), such as alternation, date of birth or first letter of name, are not considered adequate because of their predictability, which allows for the scheduling of participants, and the possible correlation between the item used (e.g. month of birth, first letter of name) and the outcome.

When implementing the sequence of randomization, if the allocation is not concealed before the patient is assigned to a group, all benefits of randomization are lost. Unconcealed randomization can lead to clinicians scheduling patients so that patients with particular characteristics receive a certain allocation, thereby biasing the allocation. Kunz et al. reviewed all available evidence of the effect of lack of concealment of allocation and concluded that studies with inadequate allocation concealment resulted in 35--40% larger estimates of treatment effect.5

In the case of indistinguishable treatment and placebo (e.g. same appearance, same schedule of administration, same taste), the care provider and the patient are blinded and the allocation concealment is self-evident: pre-numbered drugs are administered serially by the care provider. In other cases, some approaches that assure adequate concealment schemes are as follows:

<BLF>Centralized randomization (e.g. after patient consent is obtained, the investigator calls a 24-hour free-phone service to obtain the allocation group) or pharmacy-controlled randomization.

<BL>Pre-numbered or coded identical, sealed containers administered serially to participants.

<BL>On-site computer system combined with group assignments in a locked unreadable computer file that can be assessed only after entering the characteristics of an enrolled subject.

<BLL>Sequentially numbered, opaque, sealed envelopes: however, this method has to be monitored diligently to ensure that investigators do not open several envelopes beforehand and then allocate patients to the desired treatment.

Proper randomization increases the probability of obtaining similar groups at baseline and thus prevents selection bias. However, some differences in baseline characteristics between groups may appear because of chance. Important demographic and clinical characteristics for each study group should be described so that readers can assess how comparable the groups were at baseline for the known prognostic factors. The magnitude and direction of the differences are more important than results of significance statistical tests in detecting potential subversion of the randomization. Several large differences all favouring one group would heighten suspicion about the results of the trial.

<B>Were patients, care providers and outcome assessors blinded?

Blinding refers to keeping key people, such as patients, health-care providers (i.e. those administering the treatment), and outcome assessors (i.e. those assessing the main outcome), unaware of the treatment administered.

Although the term ‘double blind’ implies that neither the care provider nor the patient knows which treatment was received, it is ambiguous in terms of the blinding of other people, including those assessing patient outcomes.6 Authors should state and readers should carefully assess who was blinded (patients, care providers, outcome assessors, or monitors).

Blinding of patients and health-care providers prevents performance bias. This bias can occur if additional therapeutic interventions (i.e. co-interventions) are provided preferentially in one of the comparison groups. Blinding guarantees the same follow-up, the same attention to the patient, and the same ‘placebo effect’ in the two groups.

Blinding of outcome assessors minimizes the risk of detection bias. This type of bias occurs if patient assignment influences the process of outcome assessment, and it thus depends on both the blinding status of the assessor and the nature of the outcome. Blinding of outcome assessors is particularly important for subjective outcomes (e.g.

number of inflammatory joints), which entail increased opportunity for bias in measurement.7 Objective (hard) outcomes leave less opportunity, if any, for detection bias. Knowledge of the intervention would not greatly affect measurement of a hard outcome such as death,8 although it might influence subjective assessment such as death from cardiovascular causes. For patient-reported outcomes (PRO) such as pain, blinding of the outcome assessor implies blinding of the patient. When the two treatments under study are indistinguishable (e.g. same characteristics, same schedule of administration, same dosage), the blinding of patients, care providers and outcome assessors is easy to achieve. If treatments differ, a 'double-dummy' procedure may be useful but is not always feasible. In a double-dummy procedure, the patients of group A receive drug A and a placebo of drug B and patients of group B receive drug B and a placebo of drug A.

In other situations, clinicians and patients cannot be blinded (e.g. surgery, regimen, rehabilitation, or psychotherapy) or blinding seems feasible but cannot be effective (for specificity of adverse effects, such as bradycardia with beta-blockers). In these situations, methods to blind outcome assessors are particularly useful to avoid detection bias.9 These methods rely mainly on a centralized assessment of the main outcome, which is easy to implement for clinical investigations (e.g. laboratory tests or radiography) or clinical events (blinded adjudication committee) but requires more inventive solutions for physician-driven data (such as videotaping, audiotaping or photography of clinical examination).9,10

<B>Was follow-up complete and were patients analysed in the groups to which they were randomized?

If a rigorous trial has been undertaken to avoid all the biases that can affect the results of the study, an incorrect analytical approach can also introduce bias. In a randomized superiority trial, the most robust analytical method that prevents attrition (exclusion) bias is the intention-to-treat analysis (ITT).11 An ITT analysis means that all patients are analysed in the group to which they were initially randomized, even if they 'cross over' to the other intervention arm, they discontinue the intervention, or they are lost to follow-up. This analysis is of particular importance because participants who do not comply with the allocated treatment usually do not have the same prognostic factors as those who

comply. For per-protocol or on-treatment analysis, the analysis is restricted to participants who fulfil the protocol in terms of eligibility, interventions (treatment received) and outcome assessment. In a treatment-received analysis, patients are analysed according to the treatment they actually received, regardless of the treatment they were originally allocated to receive. Only an ITT analysis ensures that the balance in prognostic factors arising from the randomization is maintained.12 An ITT analysis answers the question 'Which choice of treatment is better for the patient?' and not 'Which treatment received is better for the patient?'. Only the former question can be answered without bias and, moreover, is the most pragmatic in choosing a treatment. Thus, ITT analysis should be the analysis of choice.

Many authors claim they performed an ITT analysis when in fact they have not: patients are excluded from the analysis if they never received any treatment, they were randomized but ineligible for the study, or they were lost to follow-up or the outcome was not assessed (e.g. arthroscopy not realized). One must check the ITT assumption by looking at the flow chart of the progress of patients through the phases of the trial and comparing the number of patients randomized to the number analysed (Figure 2). For example, in a study comparing three kinds of mattress, patients with low back pain who were randomized but who never used their mattress were not included in analysis. The conclusion was in favour of the water-bed but the risk of attrition bias is high because the rate of dropout was four times as much higher in the water-bed arm than in the two others arms.13

<Figure 2 here>

The only acceptable exclusions from an analysis are, in a strictly double-blinded study, patients who did not receive any treatment; this is usually called modified or quasi ITT analysis. Because patients do not know the treatment they will receive, their exclusion is unlikely to be due to disillusion in the allocation. If the attrition rate is low and is equally distributed between the study arms, the analysis is unlikely to be too biased.

Performing an ITT analysis usually implies choosing a method to handle missing data. Because missing data may occur for various reasons, including adverse events related to

the treatment, the method used for data imputation must be conservative, that is, not favour the treatment group.14

Some studies also report in addition efficacy data for patients willing to continue experimental treatment in follow-up extension studies. These results must be interpreted with caution because they are of course not analysed under the ITT principle.

If a large proportion of patients cross over to the opposite treatment arm or are lost to follow-up, the interpretation of study results is difficult, and neither an ITT nor a per-protocol analysis would provide reliable information. An extreme example is the Spine Patient Outcomes Research Trial (SPORT) trial, which compares standard open diskectomy and non-operative treatment for patients with lumbar intervertebral disk herniation; only 60% of patients assigned to surgery received surgery, whereas 45% of those assigned to non-operative treatment received surgery.15 Whatever the analysis performed, none will be informative.

<B>Was the outcome appropriate?

A crucial issue in assessing the results of trials is the actual measure of whether the treatment works. The outcome chosen to conclude the effectiveness of the treatment could be a clinical event (death, fracture), a therapeutic decision (length of stay, transfusion, surgery), a patient-reported outcome (pain) or a result of a complementary test (biological or morphological). Most outcomes can be measured as dichotomous variables (e.g. event/no event), as continuous variables (e.g. blood pressure, glycaemia, Western Ontario and MacMaster Universities Osteoarthritis Index score) or as time to the onset of an event (survival time data). Whatever the nature, a good outcome must have the following qualities:

<BLF>clinical relevance

<BL> good reliability and reproducibility

<BL>uniqueness

<BLL>availability for all patients to avoid attrition bias.

In addition, readers should be sceptical of studies involving unconventional outcomes not recognized in other studies.

<B>Clinical relevance and surrogate outcomes

To decide whether to apply the study results in clinical practice, one needs evidence that the treatment studied in an RCT improves outcomes that are important to patients. An elevated blood pressure outcome is of minor consequence to the patient, whereas a stroke is of major importance.

Good outcomes that are clinically relevant for the patient are death, length of hospital stay, myocardial infarction, fractures, and quality of life. The length of follow-up has to be consistent with disease evolution. For example, a follow-up of only 1 month is meaningless for a chronic disease. Kyriakidi et al. showed that only 11% of RCTs of systemic sclerosis had a follow-up of more than 1 year. 16

However, these outcomes are usually substituted by ‘surrogate’ outcomes, usually biological or imaging markers, which are easier to measure and believed to be indirect measures of the clinically relevant outcome. For example, change in bone mineral density is often used as a surrogate outcome to measure the effectiveness of treatments for the prevention of osteoporotic fractures.

As well as being of questionable clinical relevance, surrogate outcomes are often misleading.17 The effectiveness of the use of sodium fluoride is a good example of a misleading conclusion: the treatment substantially increases bone mineral density but does not prevent fractures.18

Surrogate outcomes are widely used because observing a difference in a surrogate measure requires a much smaller sample size and shorter follow-up than a clinical outcome. Trials designed to observe changes in bone mineral density require only a few hundred participants, whereas those designed to observe a fracture endpoint require many thousands of participants. However, surrogate outcomes are useful in helping guide research at its earliest stages.

<B>Reliability and reproducibility

A good outcome must be relevant for the patient but also easy to assess. Reliability of a study is affected by the reproducibility of the outcome used. For example, assessment of joint-space narrowing is more reliable if performed by two trained, independent observers and if technical acquisition of radiography is standardized. A measure of reproducibility, corrected for agreement by chance, such as the kappa coefficient, helps in assessing the quality of the measure: a value > 0.6 implies good agreement and a value > 0.8 very good agreement. Where available and appropriate, previously developed and validated scales or outcomes should be used, both to enhance quality of measurement and to assist in comparison with similar studies. Authors should indicate the origin and properties of scales.

<B>Unique primary outcome

A single primary outcome must be defined *a priori*. The study must be designed and the sample size calculated to demonstrate whether the treatment has an effect or not on this primary outcome. The rationale for this procedure is to avoid a multiplicity of statistical tests that can lead to erroneous conclusions because of the risk of hazard. Multiple analyses of the same data incur considerable risk of false-positive findings.

The alpha level is the chance taken by researchers to make a type I error: incorrectly declaring a difference to be true because of only chance producing the observed state of events. Customarily, the alpha level is set at 0.05, that is, in no more than one in 20 statistical tests will the test show some effect when in fact no effect exists. If more than one statistical test is used, the chance increases of finding at least one test result in the whole experiment that is statistically significant due to only chance and to incorrectly declare a difference or relationship to be true. In five tests, this chance is 22%; in ten, the chance increases to 40%.

For the same reason, when outcomes are assessed at several time points after randomization, the time point of primary interest must also be defined *a priori*. Many trials recruit participants over a long period. If an intervention is working particularly well or badly, the study may need to be ended early for ethical reasons. Interim analysis

could be performed. To not bias the overall study results, the interim analysis must be planned in advance, statistical methods adapted and results of analysis interpreted by an independent committee who may decide or not to stop the study.

Despite these recommendations, in 2004, Chan et al. found that more than 60% of trials had at least one primary outcome that was changed, introduced or omitted between when the protocol was approved by a scientific-ethics committee and the publication of the results. 19 Protocols, with pre-planned primary outcomes, are now publicly available on clinical research registries (e.g. www.clinicaltrials.gov, www.controlled-trials.com/isrctn) to enable the identification of outcome reporting bias.

Other outcomes of interest are secondary outcomes. All secondary outcomes must also be pre-specified and reported, not just those showing a statistically significant difference between groups. Important outcomes must be considered, but a single study must not have too many outcomes.

<B>Were results obtained from subgroup analysis?

Multiple analyses of the same data incur considerable risk of false-positive findings. As previously discussed, this risk should lead to a limitation of the number of outcomes and the number of occasions when they are assessed. The same risk is involved in multiple analyses of the same outcome in different subgroups of patients. Because of the high risk of spurious findings, subgroup analyses do not have good credibility. When considering results of a subgroup analysis, the following must be remembered:20

<BLF>Subgroup analysis should be pre-planned and should be limited to a small number of clinically important questions.

<BL>If important subgroup-treatment-effect interactions are anticipated, trials should ideally be powered to detect such interactions reliably.

<BL>Significance of the effect of treatment in individual subgroups should not be reported; rates of false-negative and false-positive results are extremely high. The only

reliable statistical approach is to test for a subgroup-treatment effect interaction, but few trials are sufficiently powered to detect this.

<BL>All subgroup analyses that were done should be reported.

<BLL>The best test of validity of subgroup-treatment effect interactions is their reproducibility in other trials.

<A>How do I understand the results?

First, for each outcome, study results should report a summary of the outcome for each group (e.g. the proportion of participants with the event, or the mean and standard deviation [SD] of measurements).2 Then, to appreciate the treatment effect, two additional data must be reported:

<NL>1. The contrast between the two groups, known as the estimation of treatment effect.

<NL>2. The precision of this estimation, the statistical significance of the treatment effect [confidence interval (CI) and/or P value].

<B>Estimation of treatment effect

For a dichotomous outcome, the estimation of treatment effect could be the risk ratio [relative risk (RR)], or risk difference [absolute risk reduction (ARR)]; for survival-time data, the measure could be the hazard ratio or difference in mean survival time; for continuous data, the estimation is usually the difference in means. The *P* value is useless for assessing the size of treatment effect.

<B>Dichotomous outcome

Consider, for example, a study in which 20% of patients of a control group died, but only 15% of patients receiving a new treatment died. The most common way to express the impact of treatment would be the RR: the risk of events for patients receiving the new treatment relative to the risk for patients in the control group:

<EQN>RR = 0.15/0.20 = 0.75 (= 75%)

An RR of 0.75 means that the new treatment reduced the risk of death in the treated group by 25% (=1 – 0.75) as compared with that in the control group. The closer the RR is to 1, the less effective the therapy. In survival analysis, RR is usually computed over a period of time and called a hazard ratio. In some statistical computations, particularly covariate adjustment, the odds ratio (OR) is computed instead of the RR. The OR is the ratio of events to non-events in the intervention group over the ratio of events to non-events in the control group. An OR can reasonably be interpreted as a RR as long as the outcome event is rare.

For some treatments and conditions, the benefit of a specific treatment, as measured by the RR, remains approximately constant over patient populations at varying baseline risk.20 As a single estimate of treatment effect can be provided for a broad class of patients, RR appears attractive. However, it is often clinically important to consider the baseline (control) risk of an event before recommending treatment because, for a given RR, the expected absolute benefit of treatment could vary considerably as the baseline risk changes. For example, an estimated RR of 50% might be important for patients at moderate to high risk of a particular adverse event. However, for patients with a low probability of an event, the risk reduction might not be sufficient to warrant the toxic effects and cost of treatment.21

Thus, the absolute risk reduction (ARR) could be considered a better measure of treatment effect because it reflects the expected absolute benefit, taking into account the baseline risk of the patient. The ARR is the difference between the proportion of control patients who die and the proportion of treatment patients who die. In our example, the ARR is as follows:

<EQN>ARR = 0.20 – 0.15 = 0.05 = 5%

For the ARR to be meaningful for clinical practice, its reciprocal, the number needed to treat (NNT) is usually used.22 The NNT is the number of patients who would need to be treated with the new treatment rather than the standard treatment for one additional

patient to benefit.23 The NNT can be obtained for any trial with a dichotomous outcome.The NNT is simply the reciprocal of the ARR:

<EQN>NNT = 1/ARR

In our example, the NNT would be 1/0.05 = 20. To prevent one additional death, 20 patients would need the treatment.

A large treatment effect, in the absolute scale, leads to a small number needed to treat. A treatment that leads to one saved life for every 10 patients treated is clearly better than a competing treatment that saves one life for every 50 treated. A correctly specified NNT must always give the comparator, the therapeutic outcome, the duration of treatment necessary to achieve that outcome, the 95% CI and the baseline risk of event without treatment.

<B>Continuous outcome

For continuous outcomes, the measure of treatment effect is the difference in means between the treatment and control group. The SD reflects the dispersion of values around the mean.

Results for continuous outcomes are usually difficult for clinicians to use in clinical practice because of problems in assessing the clinical importance of such outcomes or in comparing benefits and risks across various therapeutic options. Dividing the difference in means by the SD allows for easier comparison of effects across studies. This ratio is called the effect size (ES). For example if the mean score ± SD on a 0--100-mm visual analogue scale is 6 ± 2.5 mm in the treatment group and 5.3 ± 2.1 mm in the control group, the ES is as follows:

<EQN>ES = (Mean $_{\text{treatment}}$ – Mean $_{\text{control}}$) / SD_{pooled} = (6 – 5.3)/2.3 = 0.3

Opinions vary on how to interpret effect size, but approximately 0.2 is an indication of a small effect, 0.5 a medium effect and 0.8 a large effect size.

Translating a continuous measure (e.g. health assessment questionnaire score, visual analogic scale) to a dichotomous measure such as ‘therapeutic success (yes/no)’ usually leads to more clinically meaningful results. This dichotomization can help in interpreting and assessing the clinical significance of trial results. It allows for an NNT calculation that is usually more meaningful than the difference in means. Clinicians would find the statement ‘one needs to treat five patients to halve the EVA score in 1 patient’ more meaningful than ‘the difference in EVA score between treatment and placebo is 9.7 mm (SD = 3.7)’. However, dichotomizing a continuous variable can result in loss of information and statistical power.

<B>Precision of the estimation

<C>Confidence interval

Realistically, the true measure of the treatment effect can never be known. The best we have is the estimate provided by rigorous controlled trials. This estimate is called a point estimate, a single value calculated from observations of the sample. We usually use the 95% CI to estimate the neighbourhood within which the true effect likely lies; that is, the range that includes the true value of the effect 95% of the time. For example, if a trial involved 100 patients randomized to the treatment group and 100 to the control group, and 15 deaths occurred in the treatment group and 20 in the control group, the authors would calculate a point estimate for the RR of 0.75. However, the true RR might be much smaller or much greater than this 0.75 with a difference of only five deaths. In fact, the treatment might provide no benefit (an RR of 1) or might even do harm (RR > 1). And these suppositions would be right. In fact, these results are consistent with both an RR of 1.38 (i.e. patients given the new treatment might be 38% more likely to die than control patients) and an RR of 0.41 (i.e. patients subsequently receiving the new treatment might have a risk of dying almost 60% less than that of non-treated patients). In other words, the 95% CI for this RR is (0.41 – 1.38), and one cannot conclude that the treatment is better. If the trial enrolled 1000 patients per group, and the same event rates were observed, the point estimate of the RR would still be 0.75, but the CI would be (0.59 – 0.91), and then one could conclude that the treatment is better.

What these examples show is that the larger the sample size of a trial, the larger the number of outcome events, and the greater our confidence that the true RR (or any other measure of efficacy) is close to what we have observed. The point estimate -- in this case 0.75 -- is the one value most likely to represent the true RR. Values farther from the point estimate become less consistent with the observed RR.

When the CI of a ratio contains 1 (or 0 for a difference) the difference is not statistically significant and the result is compatible with no effect.

<C>*P* value

Many journals require or strongly encourage the use of CIs, and results should not be reported solely as *P* values.3 Yet CIs are not always reported, and one needs to know how to interpret *P* values.

Depending on the test used, there are many ways to calculate a *P* value, but its meaning is always the same. The *P* value shows how often the results would have occurred by chance if no difference existed between the two groups. In other words, the *P* value describes the risk of a false-positive conclusion of a difference when, in truth, no difference exists.

The *P* value reflects the statistical significance of a difference but not its size. A small difference observed with a large sample is more significant statistically than the same difference observed with a small sample. Thus, a difference could be statistically significant but clinically insignificant. The *P* value tells only if the observed difference is likely to be true ($P < 0.05$) or only the result of chance ($P > 0.05$), that is, statistically not significant. A $P < 0.05$ means that the result would have arisen by chance in less than one occasion in 20; a $P > 0.05$ means that the CI of the RR does not contain 1.

<B>Clinical significance vs. statistical significance

When looking at the results of a study, one must consider two important concepts: clinical significance and statistical significance. The former addresses the size of the treatment effect and the later its credibility.

The literature is full of statistically significant but clinically insignificant results. A clinically significant finding would be one clinically useful for a patient. For example, a study might find that a certain medication causes a statistically significant ($P < 0.05$) decrease in blood pressure of 2 mmHg, but this decrease would not be clinically significant. By contrast, a finding of a statistically significant decrease in blood pressure of 20 mmHg would be more clinically significant.

Statistical significance depends on the size of the difference between the groups and on the number of patients. The *P* value alone gives no information on the magnitude of the effect. Clinically trivial differences can be statistically significant if the sample size is sufficiently large. Conversely, clinically important differences can be statistically non-significant if the sample size is too small, that is, if the study lacks power (Figure 3). As an example, Keen et al. found that of 50% of rheumatology reports of RCTs with negative or indeterminate results published in 2001--2002, the studies were underpowered.24

<Figure 3 here>

Sample-size calculation for dichotomous outcomes requires four components: type I error (alpha), power, event rate in the control group and a minimal treatment effect of interest (or, analogously, an event rate in the treatment group). Calculation of continuous outcomes requires, instead of event rate in the control and treatment groups, difference a between means and assumptions on the SD.

The sample size is probably too small if in the calculation one of the following is present:

<BLF>the clinically relevant minimal treatment effect assumed is too large (a smaller but still clinically relevant difference should have been chosen)

<BL>the event rate in the control group is overestimated

<BLL>the SD is underestimated for continuous outcomes.

These over-optimistic assumptions are common because sample size calculation is often driven by feasibility. As stated by Guyatt et al., ‘investigators typically decide how many

patients they can feasibly enrol and then find ways of making assumptions that will justify embarking on a trial with a feasible sample size'.25 However, underpowered trials are not useless because their results contribute to the body of knowledge and are useful for meta-analysis. It is difficult to draw conclusions from a single trial, even large ones, but provided that the trial is not methodologically biased, the results contribute to the larger body of evidence.

<B>Independence

Standard methods of analysis assume that the data are 'independent'. For RCTs, this independence usually means that each comparison test involves only one observation per participant. Treating multiple observations from one participant as independent data is a serious error; such data arise when outcomes can be measured at successive times or from different parts of the body, as in rheumatology. For example, in a trial of osteoporosis, treating two vertebral fractures in the same patient as two independent observations is incorrect. The correct approach is to count the number of patients with at least one vertebral facture. Data analysis should be based on counting each participant once,26,27 or should involve specific statistical procedures taking into account paired data.

<A>To whom do results apply?

<B>External validity

The section 'Can I trust the results?' explored only internal validity, that is, the validity of the results in the context of the study. If the result is considered reasonably valid, next we need to explore the external validity of the result, that is, its validity in other contexts.2 Of course, if the study result is not valid, even for the subjects studied, its applicability to other groups of subjects is irrelevant.

There is no external validity *per se*. The results of a study will never be relevant to all patients and all settings, but studies should be designed and their results reported so that clinicians can judge to whom the results can reasonably be applied.

The following criteria must be considered before applying results to patients:28

<BLF>Setting of the trial: country; health-care system; recruitment from primary, secondary or tertiary care; selection of participating centres and clinicians.

<BL>Selection of patients: eligibility and exclusion criteria, ‘run-in’ or ‘washout’ period, ‘enrichment’ strategies, ratio of randomized patients to eligible non-randomized patients.

<BL>Characteristics of randomized patients: severity or stage in the natural history of disease, co-morbidities, racial group, other baseline clinical characteristics.

<BL>Difference between trial protocol and routine practice: relevance of control intervention, co-interventions, prohibition of certain non-trial treatments, therapeutic or diagnostic advances since the trial was performed.

<BL>Outcome measure and follow-up: clinical relevance of outcomes (e.g. surrogate, complex scale, composite outcome), frequency of follow-up, adequacy of the length of follow-up.

<BLL>Adverse effect of treatment: completeness of reporting of adverse effects, rate of discontinuation of treatment, selection of trial centres and/or clinicians on the basis of skill or experience, exclusion of patients at risk for complications, exclusion of patients who experienced adverse effects during a run-in period, intensity of trial safety procedure.

The clinical setting is never exactly the same as the trial setting and the patient often has attributes or characteristics different from those enrolled in the trial. These differences can result in less benefit, as was shown between a Phase III clinical trial and a Phase IV cohort study of the use of etanercept for rheumatoid arthritis.29 So one must ask whether these differences might really diminish the treatment response or greatly increase the risk of adverse events, that is, ‘Is my patient so different from the study patients that I cannot apply the results to my patient?’

<B>What are the adverse effects?

Before applying results of a study to a patient, one must consider the possible harm that any intervention might do: *primum non nocere*. Regardless of this crucial information, reporting of harms from RCTs has received less attention than reporting of efficacy, and is often inadequate. In 2004, an extension of the CONSORT statement focused on better reporting of harms-related data from RCTs.30

Computing the number needed to harm (NNH) and comparing it with the NNT to weigh benefits and risks can help to evaluate the usefulness of a treatment. The NNH is the average number of subjects receiving treatment that would lead to one additional subject having a given adverse event, as compared with the control intervention. The calculation of NNH is similar to that of the NNT:

<EQN>NNH = 1/(proportion of adverse events in the treatment group – proportion of adverse events in the control group)

The following example illustrates how to use NNT and NNH to weigh the benefits and risks of a new treatment, according to the characteristics of patients. The results of clinical trials suggest that hormone replacement therapy reduces the RR of spine fracture over a lifetime by approximately 30%, but such therapy also increases the risk of stroke by 50%. Consider two menopausal women with different baseline expected rates of spine fracture and stroke: patient A has low bone mineral density but no cardiovascular risk factors; patient B has normal bone mineral density but many cardiovascular risk factors. Table 1 summarizes the NNT and NNH values for these two women. Treating approximately 200 women such as patient A during 2 years will prevent twelve spine fractures (200/17) but induce one stroke. Given the small increased risk of stroke and the number of spine fractures prevented, many clinicians might suggest hormone replacement therapy for such patients. By contrast, treating approximately 200 women such as patient B during 2 years will prevent only six spine fractures (200/33) but will induce three strokes. Obviously, hormone replacement therapy is less indicated in these patients.

<Table 1 here>

Looking at the rate of adverse effects reported in each group helps when appraising the risks of a treatment. However, studies seldom, if ever, have enough power to detect a statistically significant difference between groups in these rates. Sample sizes are usually too small for a reasonable chance of detecting an unexpected adverse effect. Therefore, some authors suggest the creation of a composite outcomes basket to be used as the primary safety outcome in clinical effectiveness trials.31

<B>Conflict of interest

Last, but not least, another piece of information of importance when reading an article of the results of an RCT is the study funding. An article reporting industry-supported study results could be less objective than one reporting an academic-supported study. For example, in a review of results of trials comparing non-steroidal anti-inflammatory drugs (NSAIDs) used in the treatment of arthritis, Rochon et al. showed that the manufacturer-associated NSAID was almost always reported as being equal or superior in efficacy and lack of toxic effects to the comparison NSAID.32

<PP>Practice points

<BLF>Patient assignment to groups of treatment must be characterized by unpredictability: allocation randomized and concealed until patients are assigned to a group.

<BL>Blinding is the best way to avoid performance and detection bias. Studies should clearly state who was blinded (care providers, patients, outcome assessors and/or data analysts), except when it is obvious than everyone is blinded (if the two treatments are really indistinguishable).

<BL>Intention-to-treat analysis is the most robust analytical method. All patients are analysed in the group to which they were initially randomized, even if they cross over to the other intervention arm, they discontinued the intervention or they are lost to follow-up.

<BL>The primary outcome must be unique, be clinically relevant, be available for all patients, have a good reliability and reproducibility and be used for sample size calculations.

<BLL>Study results should report both a measure of treatment effect and an estimation of the precision of this measure. Readers need to critically appraise the statistical and clinical significance of a result.

<A>Conflict of interest statement

The authors declare that they have no conflict of interest. The authors have no financial or personal relationship with other people or organizations that could inappropriately influence the content of this article.

<A>References

*1. Guyatt GH, Sackett DL, Cook DJ. Users' guides to the medical literature. II. How to use an article about therapy or prevention. A. Are the results of the study valid? Evidence-Based Medicine Working Group. Jama 1993. 270: 2598-601.

*2. Guyatt GH, Sackett DL, Cook DJ. Users' guides to the medical literature. II. How to use an article about therapy or prevention. B. What were the results and will they help me in caring for my patients? Evidence-Based Medicine Working Group. Jama 1994. 271: 59-63.

*3. Moher D, Schulz KF, Altman D. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials. Jama 2001. 285: 1987-91.

4. Greenhalgh T. Assessing the methodological quality of published papers. Bmj 1997. 315: 305-8.

5. Kunz R, Vist G, Oxman AD. Randomisation to protect against selection bias in healthcare trials. Cochrane Database of Systematic Reviews 2007, Issue 2. Art No.:MR000012. DOI: 10.1002/14651858.MR000012.pub2.

6. Montori VM, Bhandari M, Devereaux PJ, et al. In the dark: the reporting of blinding status in randomized controlled trials. J Clin Epidemiol 2002. 55: 787-90.

7. Wood L, Egger M, Gluud LL, et al. Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: meta-epidemiological study. Bmj 2008. 336: 601-5.

8. Schulz KF, Grimes DA. Blinding in randomised trials: hiding who got what. Lancet 2002. 359: 696-700.

*9. Boutron I, Estellat C, Guittet L, et al. Methods of blinding in reports of randomized controlled trials assessing pharmacologic treatments: a systematic review. PLoS Med 2006. 3: e425.

*10. Boutron I, Guittet L, Estellat C, et al. Reporting Methods of Blinding in Randomized Trials Assessing Nonpharmacological Treatments. PLoS Med 2007. 4: e61.

11. Hollis S, Campbell F. What is meant by intention to treat analysis? Survey of published randomised controlled trials. Bmj 1999. 319: 670-4.

12. Heritier SR, Gebski VJ, Keech AC. Inclusion of patients in clinical trial analysis: the intention-to-treat principle. Med J Aust 2003. 179: 438-40.

13. Bergholdt K, Fabricius RN, Bendix T. Better backs by better beds? Spine 2008. 33: 703-8.

14. Baron G, Boutron I, Giraudeau B, et al. Violation of the intent-to-treat principle and rate of missing data in superiority trials assessing structural outcomes in rheumatic diseases. Arthritis Rheum 2005. 52: 1858-65.

15. Weinstein JN, Tosteson TD, Lurie JD, et al. Surgical vs nonoperative treatment for lumbar disk herniation: the Spine Patient Outcomes Research Trial (SPORT): a randomized trial. Jama 2006. 296: 2441-50.

16. Kyriakidi M, Ioannidis JP. Design and quality considerations for randomized controlled trials in systemic sclerosis. Arthritis Rheum 2002. 47: 73-81.

17. Fleming TR, DeMets DL. Surrogate end points in clinical trials: are we being misled? Ann Intern Med 1996. 125: 605-13.

18. Riggs BL, Hodgson SF, O'Fallon WM, et al. Effect of fluoride treatment on the fracture rate in postmenopausal women with osteoporosis. N Engl J Med 1990. 322: 802-9.

19. Chan AW, Hrobjartsson A, Haahr MT, Gotzsche PC, Altman DG. Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles. *Jama*. 2004;291:2457-65.

*20. Rothwell PM. Treating individuals 2. Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. Lancet 2005. 365: 176-86.

*21. Cook RJ, Sackett DL. The number needed to treat: a clinically useful measure of treatment effect. Bmj 1995. 310: 452-4.

22. Osiri M, Suarez-Almazor ME, Wells GA, et al. Number needed to treat (NNT): implication in rheumatology clinical practice. Ann Rheum Dis 2003. 62: 316-21.

23. Laupacis A, Sackett DL, Roberts RS. An assessment of clinically useful measures of the consequences of treatment. N Engl J Med 1988. 318: 1728-33.

24. Keen HI, Pile K, Hill CL. The prevalence of underpowered randomized clinical trials in rheumatology. J Rheumatol 2005. 32: 2083-8.

25. Guyatt GH, Mills EJ, Elbourne D. In the era of systematic reviews, does the size of an individual trial still matter. PLoS Med 2008. 5: e4.

26. Altman DG, Bland JM. Statistics notes. Units of analysis. Bmj 1997. 314: 1874.

27. Bolton S. Independence and statistical inference in clinical trial designs: a tutorial review. J Clin Pharmacol 1998. 38: 408-12.

*28. Rothwell PM. External validity of randomised controlled trials: 'to whom do the results of this trial apply?' Lancet 2005. 365: 82-93.

29. Farahani P, Levine M, Gaebel K, et al. Clinical data gap between phase III clinical trials (pre-marketing) and phase IV (post-marketing) studies: evaluation of etanercept in rheumatoid arthritis. Can J Clin Pharmacol 2005. 12: e254-63.

30. Ioannidis JP, Evans SJ, Gotzsche PC, et al. Better reporting of harms in randomized trials: an extension of the CONSORT statement. Ann Intern Med 2004. 141: 781-8.

31. Tugwell P, Judd MG, Fries JF, et al. Powering our way to the elusive side effect: a composite outcome 'basket' of predefined designated endpoints in each organ system should be included in all controlled trials. J Clin Epidemiol 2005. 58: 785-90.

32. Rochon PA, Gurwitz JH, Simms RW, et al. A study of manufacturer-supported trials of nonsteroidal anti-inflammatory drugs in the treatment of arthritis. Arch Intern Med 1994. 154: 157-63.

Table 1 Risk of spine fracture or stroke with and without HRT, relative risk, absolute risk reduction or increase and NNT or NNH for two profiles of patient. Patient A has low bone mineral density but no cardiovascular risk factors; patient B has normal bone mineral density but many cardiovascular risk factors

		Risk without HRT	Risk with HRT	Relative risk	Absolute risk variation (reduction or increase)	NNT or NNH
Patient A	Spine fracture	0.20	0.14	0.70	0.06	17
	Stroke	0.01	0.015	1.5	0.005	200
Patient B	Spine fracture	0.10	0.07	0.70	0.03	33
	Stroke	0.03	0.045	1.5	0.015	67

<TFN>HRT: hormone replacement therapy; NNT: number needed to treat; NNH: number needed to harm

Figure 1 Sources of bias in randomised controlled trials and methods to overcome them.

Figure 2 Flow chart of participants through each stage of a randomised controlled trial.

Figure 3 Clinical significance and interpretation of the *P* value.